

Learning Semantic Segmentation with Weakly-Annotated Videos

Pavel Tokmakov*

Karteek Alahari*

Cordelia Schmid*

Inria

Abstract

Fully convolutional neural networks (FCNNs) trained on a large number of images with strong pixel-level annotations have become the new state of the art for the semantic segmentation task. While there have been recent attempts to learn FCNNs from image-level weak annotations, they need additional constraints, such as the size of an object, to obtain reasonable performance. To address this issue, we present motion-CNN (M-CNN), a novel FCNN framework which incorporates motion cues and is learned from video-level weak annotations. Our learning scheme to train the network uses motion segments as soft constraints, thereby handling noisy motion information. When trained on weakly-annotated videos, our method outperforms the state-of-the-art approach [28] on the PASCAL VOC 2012 image segmentation benchmark. We also demonstrate that the performance of M-CNN learned with 150 weak video annotations is on par with state-of-the-art weakly-supervised methods trained with thousands of images. Finally, M-CNN substantially outperforms recent approaches in a related task of video co-localization on the YouTube-Objects dataset. This is an extended version of our ECCV paper [39].

1. Introduction

The need for weakly-supervised learning for semantic segmentation has been highlighted recently [16, 32, 40]. It is particularly important, as acquiring a training set by labeling images manually at the pixel level is significantly more expensive than assigning class labels at the image level. Recent segmentation approaches have used weak annotations in several forms: bounding boxes around objects [26, 41], image labels denoting the presence of a category [32, 40] or a combination of the two [28]. All these previous approaches only use annotation in images, i.e., bounding boxes, image tags, as a weak form of supervision. Naturally, additional cues would come in handy to address this challenging problem. As noted in [5], motion is one

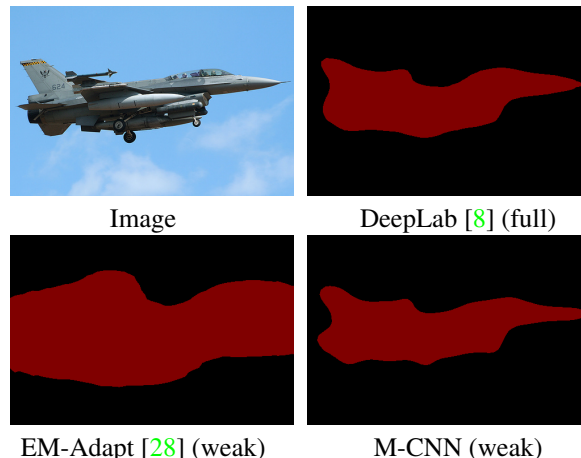


Figure 1. Comparison of state-of-the-art fully [8] and weakly [28] supervised methods with our weakly-supervised M-CNN model.

such cue for semantic segmentation, which helps us identify the extent of objects and their boundaries in the scene more accurately. To our knowledge, motion has not yet been leveraged for weakly-supervised semantic segmentation. In this work, we aim to fill this gap by learning an accurate segmentation model with the help of motion cues extracted from weakly-annotated videos.

Our proposed framework is based on fully convolutional neural networks (FCNNs) [8, 14, 25, 43], which extend deep CNNs, and are able to classify every pixel in an input image in a single forward pass. While FCNNs show state-of-the-art results on segmentation benchmark datasets, they require thousands of pixel-level annotated images to train on—a requirement that limits their utility. Recently, there have been some attempts [28, 30–32] to train FCNNs with weakly-annotated images, but they remain inferior in performance to their fully-supervised equivalents (see Fig. 1). In this paper, we develop a new CNN variant named M-CNN, which leverages motion cues in weakly-labeled videos, in the form of unsupervised motion segmentation, e.g., [29]. It builds on the architecture of FCNN by adding a motion segmentation based label inference step, as shown in Fig. 2. In other words, predictions from the FCNN

*Thoth team, Inria, Laboratoire Jean Kuntzmann, Grenoble, France.

layers and motion segmentation jointly determine the loss used to learn the network (see §3.3).

Our approach uses unsupervised motion segmentation from real-world videos, such as the YouTube-Objects [33] and the ImageNet-VID [17] datasets, to train the network. In this context, we are confronted with two main challenges. The first one is that even the best-performing algorithms cannot produce good motion segmentations consistently, and the second one is the ambiguity of video-level annotations, which cannot guarantee the presence of object in all the frames. We develop a novel scheme to address these challenges automatically without any manual annotations, apart from the labels assigned at the video level, denoting the presence of objects somewhere in the video. To this end, we use motion segmentations as soft constraints in the learning process, and also fine-tune our network with a small number of video shots to refine it.

We evaluated the proposed method on two related problems: semantic segmentation and video co-localization. When trained on weakly-annotated videos, M-CNN outperforms state-of-the-art EM-Adapt [28] significantly, on the PASCAL VOC 2012 image segmentation benchmark [13]. Furthermore, our trained model, despite using only 150 video labels, achieves performance similar to EM-Adapt trained on more than 10,000 VOC image labels. Augmenting our training set with 1,000 VOC images results in a further gain, achieving the best performance on VOC 2012 test set in the weakly-supervised setting (see §4.4). On the video co-localization task, where the goal is to localize common objects in a set of videos, M-CNN substantially outperforms a recent method [21] by over 16% on the YouTube-Objects dataset.

The contributions of this work are twofold: (i) We present a novel CNN framework for segmentation that integrates motion cues in video as soft constraints. (ii) Experimental results show that our segmentation model learned from weakly-annotated videos can indeed be applied to evaluate on challenging benchmarks and achieves top performance on semantic segmentation as well as video co-localization tasks. Code for training our M-CNN iteratively is integrated in an FCNN framework in Caffe [18], and will be made available.

2. Related Work

In addition to fully-supervised segmentation approaches, such as [6, 7], several weakly-supervised methods have been proposed over the years: some of them use bounding boxes [26, 41], while others rely on image labels [40]. Traditional approaches for this task, such as [40], used a variety of hand-crafted visual features, namely, SIFT histograms, color, texture, in combination with a graphical or a parametric structured model. Such early attempts have been recently outperformed by FCNN methods, e.g., [28].

FCNN architecture [8, 14, 24, 25, 28, 30–32, 43] adapts standard CNNs [20, 22] to handle input images of any arbitrary size by treating the fully connected layers as convolutions with kernels of appropriate size. This allows them to output scores for every pixel in the image. Most of these methods [8, 14, 24, 25, 43] rely on strong pixel-level annotation to train the network.

Attempts [28, 30–32] to learn FCNNs for the weakly-supervised case use either a multiple instance learning (MIL) scheme [31, 32] or constraints on the distribution of pixel labels [28, 30] to define the loss function. For example, Pathak *et al.* [31] extend the MIL framework used for object detection [11, 36] to segmentation by treating the pixel with the highest prediction score for a category as its positive sample when computing the loss. Naturally, this approach is susceptible to standard issues suffered by MIL, like converging to the most discriminative parts of objects [11]. An alternative MIL strategy is used in [32], by introducing a soft aggregation function that translates pixel-level FCNN predictions into an image label distribution. The loss is then computed with respect to the image annotation label and backpropagated to update the network parameters. This strategy works better in practice than [31], but requires training images that contain only a single object, as well as explicit background images. Furthermore, it uses a complex post-processing step involving multi-scale segmentations when testing, which is critical to its performance.

Weakly-supervised FCNNs in [28, 30] define constraints on the predicted pixel labels. Papandreou *et al.* [28] presented an expectation maximization (EM) approach, which alternates between predicting pixel labels (E-step) and estimating FCNN parameters (M-step). Here, the label prediction step is moderated with cardinality constraints, i.e., at least 20% of the pixels in an image need to be assigned to each of the image-label categories, and at least 40% to the background. This approach was extended in [30] to include generic linear constraints on the label space, by formulating label prediction as a convex optimization problem. Both these methods showed excellent results on the PASCAL VOC 2012 dataset, but are sensitive to the linear/cardinality constraints. We address this drawback in our M-CNN framework, where motion cues act as more precise constraints. Fig. 1 shows the improvement due to these constraints. We demonstrate that FCNNs can be trained with videos, unlike all the previous methods restricted to images, and achieve the best performance using much less training data more effectively.

Weakly-supervised learning is also related to weakly-supervised learning. Methods following this recent trend [9, 10, 12, 23] are kick-started with either a small number of manually annotated examples, e.g., some fully-supervised training examples for the object detection task in [23], or au-

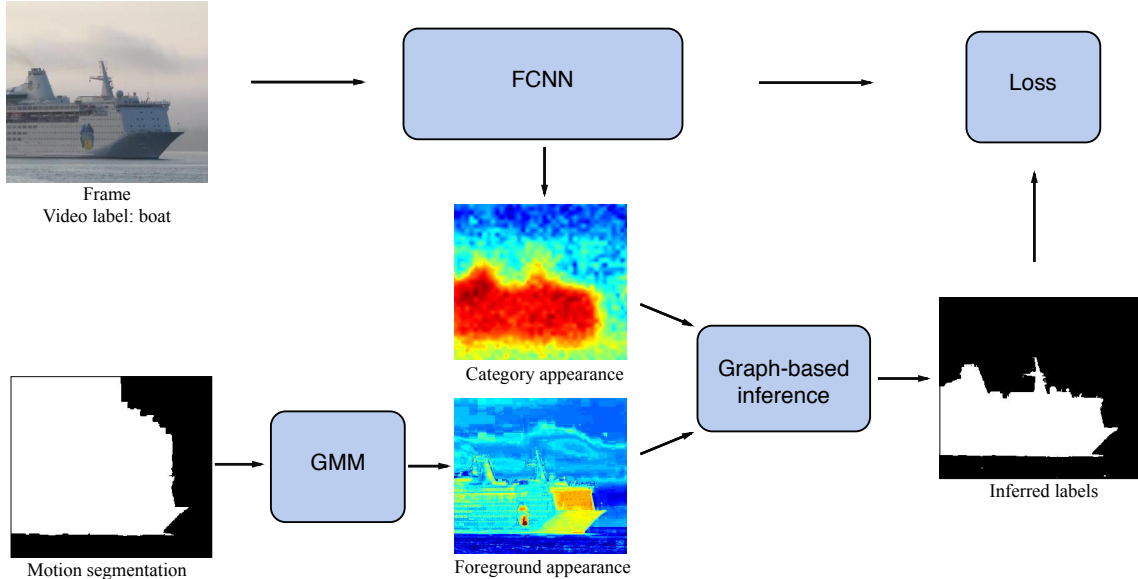


Figure 2. Overview of our M-CNN framework, where we show only one frame from a video example for clarity. The soft potentials (foreground appearance) computed from motion segmentation and the FCNN predictions (category appearance) jointly determine the latent segmentation (inferred labels) to compute the loss, and thus the network update.

tomatically discovered “easy” samples [9], and then trained with a gradually increasing set of examples mined from web resources. However, none of them address the semantic segmentation problem. Other paradigms related to weakly-supervised learning, such as co-localization [33] and co-segmentation [35] require the video (or image) to contain a dominant object class. Co-localization methods aim to localize the common object with bounding boxes, whereas in co-segmentation, the goal is to estimate pixel-wise segment labels. Such approaches, e.g., [19, 33, 38], typically rely on a pre-computed candidate set of regions (or boxes) and choose the best one with an optimization scheme. Thus, they have no end-to-end learning mechanism and are inherently limited by the quality of the candidates.

3. Learning semantic segmentation from video

We begin by presenting a summary of the entire approach in Section 3.1. We then describe the network architecture in Section 3.2, explain the estimation of latent segmentation variables and the computation of the loss function for learning the network in Section 3.3. Finally, Section 3.4 presents the fine-tuning step to further improve our M-CNN.

3.1. Overview

We train our network by exploiting motion cues from video sequences. Specifically, we extract unsupervised motion segments from video, with algorithms such as [29], and use them in combination with the weak labels at the video level to learn the network. We sample frames from all the

video sequences uniformly, and assign them the class label of the video. This collection forms our training dataset, along with their corresponding motion segments.

The parameters of M-CNN are updated with a standard mini-batch SGD, similar to other CNN approaches [28], with the gradient of a loss function. Here, the loss measures the discrepancy between the ground truth segmentation label and the label predicted at each pixel. Thus, in order to learn the network for the semantic segmentation task, we need pixel-level ground truth for all the training data. These pixel-level labels are naturally latent variables in the context of weakly-supervised learning. Now, the task is to estimate them for our weakly-labeled videos. An ideal scenario in this setting would be near-perfect motion segmentations, which can be directly used as object ground truth labels. However, in practice, not only are the segmentations far from perfect (see Fig. 3), but also fail to capture moving objects in many of the shots. This makes a direct usage of motion segmentation results suboptimal. To address this, we propose a novel scheme, where motion segments are only used as soft constraints to estimate the latent variables together with object appearance cues.

The other challenges when dealing with real-world video datasets, such as YouTube-Objects and ImageNet-VID, are related to the nature of video data itself. On one hand, not all parts of a video contain the object of interest. For instance, a video from a show reviewing boats may contain shots with the host talking about the boat, and showing it from the inside for a significant part—content that is unsuitable for learning a segmentation model for the VOC ‘boat’



Frame Motion seg. [29] Label prediction

Figure 3. Examples highlighting the importance of label prediction for handling imprecise motion segmentations (second column). The soft GMM potentials computed from motion segments together with network predictions produce better labels (third column) to learn the network. See §3.3 for details.

category. On the other hand, a long video can contain many nearly identical object examples which leads to an imbalance in the training set. We address both problems by fine-tuning our M-CNN with an automatically selected, small subset of the training data.

3.2. Network architecture

Our network is built on the DeepLab model for semantic image segmentation [8]. It is an FCNN, obtained by converting the fully-connected layers of the VGG-16 network [37] into convolutional layers. A few other changes are implemented to get a dense network output for an image at its full resolution efficiently. Our work builds on this network. We develop a more principled and effective label prediction scheme involving motion cues to estimate the latent variables, in contrast to the heuristic size constraints used in [28], which is based on DeepLab.

3.3. Estimating latent variables with label prediction

Given an image of N pixels, let \mathbf{p} denote the output of the softmax layer of the convolutional network. Then, $p_i^l \in [0, 1]$ is the prediction score of the network at pixel i for label l . The parameters of the network are updated with the gradient of the loss function, given by:

$$\mathcal{L}(\mathbf{x}, \mathbf{p}) = \sum_{i=1}^N \sum_{l=0}^L \delta(x_i - l) \log(p_i^l), \quad (1)$$

where \mathbf{x} denotes ground truth segmentation labels in the fully-supervised case, \mathbf{p} is the current network prediction, and $\delta(x_i - l)$ is the Dirac delta function, i.e., $\delta(x_i - l) = 1$, if $x_i = l$, and 0 otherwise. The segmentation label x_i of pixel i takes values from the label set $\mathbf{L} = \{0, 1, \dots, L\}$, containing the background class (0) and L object categories. Naturally, in the weakly-supervised case, ground truth segmentation labels are unavailable, and \mathbf{x} represents latent segmentation variables, which need to be estimated. We perform this estimation with soft motion segmentation cues in this paper.

Given the motion segmentation $\mathbf{s} = \{s_i | i = 1, \dots, N\}$, where $s_i \in \{0, 1\}$ denotes whether a pixel i belongs to foreground (1) or background (0).¹ The regions assigned to foreground can represent multiple object categories when the video is tagged with more than one category label. A simple way of transforming motion segmentation labels s_i into latent semantic segmentation labels x_i is with a hard assignment, i.e., $x_i = s_i$. This hard assignment is limited to videos containing a single category label, and also makes the assumption that motion segments are accurate and can be used as they are. We will see in our experiments that this performs poorly when using real-world video datasets (cf. ‘M-CNN* hard’ in Table 1). We address this by using motion cues as soft constraints for estimating the label assignment \mathbf{x} in the following.

Inference of the segmentation \mathbf{x} . We compute the pixel-level segmentation \mathbf{x} as the minimum of an energy function $E(\mathbf{x})$ defined by:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \left(\psi_i^m(z_i) + \alpha \psi_i^{fc}(p_i^{x_i}) \right) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j), \quad (2)$$

where $\mathcal{V} = \{1, 2, \dots, N\}$ is the set of all the pixels, z_i denotes the RGB color at pixel i and the set \mathcal{E} denotes all pairs of neighboring pixels in the image. Unary terms ψ_i^m and ψ_i^{fc} are computed from motion cues and current predictions of the network respectively, with α being a scalar parameter balancing their impact. The pairwise term ψ_{ij} imposes a smoothness over the label space.

The first unary term ψ_i^m captures the appearance of all foreground objects obtained from motion segments. To this end, we learn two Gaussian mixture models (GMMs), one each for foreground and background, with RGB values of pixel colors, similar to standard segmentation methods [29, 34]. The foreground GMM is learned with RGB values of all the pixels assigned to foreground in the motion segmentation. The background GMM is learned in a similar fashion with the corresponding background pixels. Given the RGB values of a pixel i , $\psi_i^m(z_i)$ is given by the negative

¹We do not include an index denoting the frame number in the video for brevity.

log-likelihood of the corresponding GMM (background one for $l = 0$ and foreground otherwise). Using motion cues to generate this soft potential ψ_i^m helps us alleviate the issue of imperfect motion segmentation. The second unary term ψ_i^{fc} represents the learned category appearance model determined by the current network prediction $p_i^{x_i}$ for pixel i , i.e., $\psi_i^{fc}(p_i^{x_i}) = -\log(p_i^{x_i})$.

The pairwise term is based on a contrast-sensitive Potts model [3, 34] as:

$$\psi_{ij}(x_i, x_j) = \lambda(1 - \Delta(i, j))(1 - \delta(x_i - x_j)) \frac{\exp(-\gamma \|z_i - z_j\|^2)}{\text{dist}(i, j)}, \quad (3)$$

where z_i and z_j are colors of pixels i and j , λ is a scalar parameter to balance the order of magnitude of the pairwise term with respect to the unary term, and γ is a scalar parameter set to 0.5 as in [29]. The function $\text{dist}(i, j)$ is the Euclidean distance between pixels. The Dirac delta function $\delta(x_i - x_j)$ ensures that the pairwise cost is only applicable when two neighboring pixels take different labels. In addition to this, we introduce the term $(1 - \Delta(i, j))$, where $\Delta(i, j) = 1$ if pixels i and j both fall in the boundary region around the motion segment, and 0 otherwise. This accounts for the fact that motion segments may not always respect color boundaries, and allows the minimization algorithm to assign different labels to neighboring pixels around motion edges.

We minimize the energy function (2) with an iterative GrabCut-like [34] approach, wherein we first apply the alpha expansion algorithm [4] to get a multi-label solution, use it to re-estimate the (background and foreground) GMMs, and then repeat the two steps for a few iterations. We highlight the importance of our label prediction technique with soft motion-cue constraints in Fig. 3. Here, the original, binary motion predictions are imprecise (bottom two rows) or incorrect (top row) in all the examples, whereas using them as soft constraints in combination with the network prediction results in a more accurate estimation of the latent segmentation variables.

3.4. Fine-tuning M-CNN

We learn an initial M-CNN model from all the videos in the dataset which have sufficient motion information (see §4.2 for implementation details). To refine this model we add a fine-tuning step, which updates the parameters of the network with a small set of unique and reliable video examples. This set is built automatically by selecting one shot from each video sequence, whose motion segment has the highest overlap (intersection over union) score with the current M-CNN prediction. The intuition behind this selection criterion is that our MCNN has already learned to discriminate categories of interest from the background, and thus, its predictions will have the highest overlap with precise motion segmentations. This model refinement leverages the

most reliable exemplars and avoids near duplicates, often occurring within one video. In Section 4.3 we demonstrate that this step is necessary for dealing with real-world non-curated video data.

4. Results and Evaluation

4.1. Experimental protocol

We trained our M-CNN in two settings. The first one is on purely video data, and the second on a combination of image and video data. We performed experiments primarily with the weakly-annotated videos in the YouTube-Objects v2.2 dataset [42]. Additionally, to demonstrate that our approach adapts to other datasets automatically, we used the ImageNet video (ImageNet-VID) dataset [17]. The weakly-annotated images to train our network jointly on image and video data were taken from the training part of the PASCAL VOC 2012 segmentation dataset [13] with their image tags only. We then evaluated variants of our method on the VOC 2012 segmentation validation and test sets.

The YouTube-Objects dataset consists of 10 classes, with 155 videos in total. Each video is annotated with one class label and is split automatically into shots, resulting in 2511 shots overall. For evaluation, one frame per shot is annotated with a bounding box in some of the shots. We use this exclusively for evaluating our video co-localization performance in Section 4.5. For experiments with ImageNet-VID, we use 795 training videos corresponding to the 10 classes in common with YouTube-Objects. ImageNet-VID has bounding box annotations produced semi-automatically for every frame in a video shot (2120 shots in total). We accumulate the labels over a shot and assign them as class labels for the entire shot. As in the case of YouTube-Objects, we only use class labels at the video level and none of the available additional annotations.

The PASCAL VOC 2012 dataset has 20 foreground object classes and a background category. It is split into 1464 training, 1449 validation and 1456 test images. For experiments dealing with the subset of 10 classes in common with YouTube-Objects (see the list in Table 1), we treat the remaining 10 from VOC as irrelevant classes. In other words, we exclude all the training/validation images which contain only the irrelevant categories. This results in 914 training and 909 validation images. In images that contain an irrelevant class together with any of the 10 classes in YouTube-Objects, we treat their corresponding pixels as background for evaluation. Some of the state-of-art methods [28, 30] use an augmented version of the VOC 2012 dataset, with over 10,000 additional training images [15]. Naturally the variants trained on this large dataset perform significantly better than those using the original VOC dataset. We do not use this augmented dataset in our work, but report state-of-the-art results due to our motion cues.

Method	FOV	bkg	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	Average
EM-Adapt	small	65.7	25.1	20.5	9.3	21.6	23.7	12.4	17.7	14.9	19.5	25.4	23.2 ± 3.0
EM-Adapt	large	69.1	12.9	14.7	9.0	12.9	15.4	5.6	9.9	7.8	15.9	23.0	17.9 ± 4.4
M-CNN*	small	83.4	30.3	35.2	13.5	11.6	36.5	22.1	19.8	22.2	5.2	13.7	26.7 ± 1.0
M-CNN*	large	84.6	35.3	44.8	24.7	21.7	44.4	26.3	26.5	27.9	10.0	22.9	33.6 ± 0.2
M-CNN* hard	large	83.6	35.3	38.6	24.0	21.2	39.6	20.2	21.3	19.2	7.9	17.9	29.9 ± 0.7
M-CNN	large	86.3	46.5	43.5	27.6	34.0	47.5	28.7	31.0	30.8	32.4	43.4	41.2 ± 1.3

Table 1. Performance of M-CNN and EM-Adapt variants, trained with YouTube-Objects, on the VOC 2012 validation set. ‘*’ denotes the M-CNN models without fine-tuning. ‘M-CNN* hard’ is the variant without the label prediction step. ‘M-CNN’ is our complete method: with fine-tuning and label prediction. We report average and standard deviation over 5 runs.

The segmentation performance of all the methods is measured as the intersection over union (IoU) score of the predicted segmentation and the ground truth. We compute IoU for each class as well as the average over all the classes, including background, following standard protocols [13, 28]. We also evaluate our segmentation results in the co-localization setting with the CorLoc measure [19, 29, 33], which is defined as the percentage of images with IoU score, between ground truth and predicted bounding boxes, more than 0.5.

4.2. Implementation details

Motion segmentation. In all our experiments we used [29], a state-of-the-art method for motion segmentation. We perform two pruning steps before training the network. First, we discard all shots with less than 20 frames ($2 \times$ the batch size of our SGD training). Second, we remove shots without relevant motion information: (i) when there are nearly no motion segments, or (ii) a significant part of the frame is assigned to foreground. We prune them out by a simple criterion based on the size of the foreground segments. We keep only the shots where the estimated foreground occupies between 2.5% and 50% of the frame area in each frame, for at least 20 contiguous frames in the shot. In cases where motion segmentation fails in the middle of a shot, but recovers later, producing several valid sequences, we keep the longest one. These two steps combined remove about a third of the shots, with 1675 and 1691 shots remaining in YouTube-Objects and ImageNet-VID respectively. We sample 10 frames uniformly from each of these remaining shots to train the network.

Training. We use a mini-batch of size 10 for SGD, where each mini-batch consists of the 10 frame samples of one shot. Our CNN learning parameters follow the setting in [28]. The initial learning rate is set to 0.001 and multiplied by 0.1 after a fixed number of iterations. We use a momentum of 0.9 and a weight decay of 0.0005. Also, the loss term $\delta(x_i - l) \log(p_i^l)$ computed for each object class l with num_l training samples, in (1), is weighted

by $\min_{j=1 \dots L} \text{num}_j / \text{num}_l$. This accounts for imbalanced number of training samples for each class in the dataset.

In the energy function (2), the parameter α , which controls the relative importance of the current network prediction and the soft motion cues, is set to 1 when training on the entire dataset. It is increased to 2 for fine-tuning, where the predictions are more reliable due to an improved network. We perform 4 iterations of the graph cut based inference algorithm, updating the GMMs at each step. The inference algorithm is either alpha expansion (for videos with multiple objects) or graph cut (when there is only one object label for the video). Following [29], we learn GMMs for a frame t with the motion segments from all the 10 frames in a batch, weighting each of them inversely according to their distance from t . The fine-tuning step is performed very selectively with the best shot for each video, where the average overlap is no less than 0.2.

A systematic evaluation on the VOC 2012 validation set confirmed that the performance is not sensitive to the number of iterations and the α parameter. The number of iterations is set as in other iterative graph cut based methods, e.g., [29]. In experiments on the VOC 2012 validation set, with the model trained on YouTube-Objects (M-CNN* in Table 1), we found that this has a marginal impact on the performance: changing the number of iterations from 1 through 5 resulted in average IoU scores 33.6, 33.1, 33.5, 33.6 and 33.9 respectively. The α parameter is set based on the intuition that the network predictions are more reliable in the fine-tuning step, where the network is already trained on the entire dataset. The performance is again not sensitive within a range of values, with only extreme cases changing IoU significantly: $\alpha = 0.5$: 24.7, 1.0: 33.8, 2.0: 34.1, 3.0: 34.3, 10.0: 23.3. In the fine-tuning step (M-CNN in Table 1), there is even less of an impact due to a better trained model: $\alpha = 0.5$: 41.4, 1.0: 41.9, 2.0: 42.3, 3.0: 42.6, 10.0: 42.2.

Code. We implemented our M-CNN in the Caffe framework [18], with the proposed label prediction step as a new layer. We will make our source code, configuration files,

Method	Dataset	bkg	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	Average
EM-Adapt	YTube	65.7	25.1	20.5	9.3	21.6	23.7	12.4	17.7	14.9	19.5	25.4	23.2 [†]
EM-Adapt	ImNet	66.1	22.8	18.7	16.9	26.7	35.7	22.4	23.6	21.4	28.4	24.3	27.9
EM-Adapt	VOC	75.5	30.5	27.4	24.1	41.8	36.8	25.5	33.3	29.3	40.0	29.7	35.8
EM-Adapt	VOC aug.	77.4	32.1	30.8	26.4	42.6	40.7	32.8	37.8	35.1	45.2	41.1	40.2
M-CNN	YTube	86.3	46.5	43.5	27.6	34.0	47.5	28.7	31.0	30.8	32.4	43.4	41.2 [†]
M-CNN	VOC+YTube	85.4	54.5	40.8	35.5	41.2	47.5	38.3	42.0	41.5	45.0	47.8	47.2 [†]
M-CNN	VOC aug.+YTube	82.5	47.8	35.3	29.6	45.6	54.6	40.3	46.6	44.8	52.2	56.6	48.7
M-CNN	ImNet	85.6	41.4	45.3	23.2	38.6	42.3	36.0	35.1	21.1	15.3	44.8	39.0
M-CNN	VOC+ImNet	85.1	53.3	46.8	32.5	33.9	37.3	40.7	32.3	34.2	40.0	45.0	43.7
M-CNN	VOC aug.+ImNet	83.1	47.6	40.3	26.4	44.1	51.1	41.7	51.0	34.9	44.6	52.7	47.0

Table 2. Performance of our M-CNN variants on the VOC 2012 validation set is shown as IoU scores. We also compare with the best variants of EM-Adapt [28] trained on YouTube-Objects (YTube), ImageNet-VID (ImNet), VOC, and augmented VOC (VOC aug.) datasets. [†] denotes the average result of 5 trained models.

and trained models available online [1], to allow the reproduction of all the reported results.

4.3. Evaluation of M-CNN

We start by evaluating the different components of our M-CNN approach and compare to the state-of-the-art EM-Adapt method, see Table 1. We train EM-Adapt and M-CNN with the pruned shots from our YouTube-Objects training set in two network settings: large and small field of view (FOV). The large FOV is 224×224 , while the small FOV is 128×128 . We learn 5 models which vary in the order of the training samples and their variations (cropping, mirroring), and report the mean score and standard deviation.

The small FOV M-CNN without the fine-tuning step achieves an IoU of 26.7%, whereas large FOV gives 33.6% on the PASCAL VOC 2012 validation set. In contrast, EM-Adapt [28] trained² on the same dataset performs poorly with large FOV. Furthermore, both the variants of EM-Adapt are lower in performance than our M-CNN, notably about 16% for large FOV. This is because EM-Adapt uses a heuristic (where background is constrained to 40% of the image area, and foreground to at least 20%) to estimate the latent segmentation labels, and fails to leverage the weak supervision in our training dataset effectively. Our observation on this failure of EM-Adapt is further supported by the analysis in [28], which notes that a large FOV network performs poorer than its small FOV counterpart when only a “small amount of supervision is leveraged”. The label prediction step (§3.3) proposed in our method leverages training data better than EM-Adapt, by optimizing an energy function involving soft motion constraints and network responses. We also evaluated the significance of using motion cues as soft constraints (M-CNN*) instead of introducing

them as hard labels (M-CNN* hard), i.e., directly using motion segmentation result as latent labels \mathbf{x} . ‘M-CNN* hard’ achieves 29.9 compared to 33.6 with soft constraints. We then take our best variant (M-CNN with large FOV) and fine-tune it, improving the performance further to 41.2%. In all the remaining experiments, we use the best variants of EM-Adapt and M-CNN.

4.4. Training on weakly-annotated videos & images

We also trained our M-CNN with weakly-annotated videos and images. To this end, we used images from the VOC 2012 training set. We added the 914 images from the VOC 2012 training set containing the 10 classes, and used only their weak annotations, i.e., image-level labels. In this setting, we first trained the network with the pruned video shots from YouTube-Objects, fine-tuned it with a subset of shots (as described in §3.4), and then performed a second fine-tuning step with these selected video shots and VOC images. To estimate the latent segmentation labels we use our optimization framework (§3.3) when the training sample is from the video dataset and the EM-Adapt label prediction step when it is from the VOC set. We can alternatively use our framework with only the network prediction component for images. For example, fine-tuning M-CNN with VOC and YouTube-Objects using the network prediction component only for VOC images (i.e., without EM-Adapt) improves the performance to 51.0 (from 47.2 in Table 2). The success of this depends on the quality of the network prediction, and it is not viable when training on classes without video data, i.e., the remaining 10 classes in VOC.

As shown in Table 2, using image data, with additional object instances, improves the IoU score from 41.2 to 47.2. In comparison, EM-Adapt re-trained for 10 classes on the original VOC 2012 achieves only 35.8. Augmenting the dataset with several additional training images [15], im-

²We used the original implementation provided by the authors to train EM-Adapt.

Method	Training data	# samples	Average	Average 10-class
<i>Strong/Full supervision</i>				
[32] + bb	VOC+ImNet	~762,500	37.0	43.8
[32] + seg	VOC+ImNet	~761,500	40.6	48.0
[28] + seg	VOC aug.	12,031	69.0	78.2
[27] (full)	VOC aug.	10,582	69.6	79.3
[43] (full)	VOC aug.+COCO	77,784	74.7	82.9
<i>Weak supervision with additional info.</i>				
[32] + sp	ImNet	~760,000	35.8	42.3
[30] + sz	VOC aug.	10,582	43.3	48.9
[30] + sz + CRF	VOC aug.	10,582	45.1	51.2
[28] + CRF	VOC aug.	12,031	39.6	45.2
<i>Weak supervision</i>				
[31]	VOC aug.	12,031	25.7	-
[30]	VOC aug.	10,582	35.6	39.5
[28]	VOC aug.	12,031	35.2	40.3
Ours	VOC+YTube	3,139	39.8	49.6
Ours	VOC+ImNet	3,155	36.9	48.0

Table 3. Evaluation on the VOC 2012 test set shown as IoU scores. We compare with several recent weakly-supervised methods: EM-Adapt [28], [31], [30], as well as methods using strong or full supervision: [32]+bb, [32]+seg, [28]+seg, [27,43], and those using additional information: [32]+sp, [30]+sz, [30]+sz+CRF, [28]+CRF.

proves it to 40.2, but this remains considerably lower than our result. M-CNN trained with ImageNet-VID achieves 39.0 (ImNet in the table), which is comparable to our result with YouTube-Objects. The performance is significantly lower for the motorbike class (15.3 vs 32.4) owing to the small number of video shots available for training. In this case, we only have 67 shots compared to 272 from YouTube-Objects. Augmenting this dataset with VOC images boosts the performance to 43.7 (VOC+ImNet). Augmenting the training set with additional images (VOC aug.) further increases the performance.

Qualitative results. Fig. 4 shows qualitative results of M-CNN (trained on VOC and YouTube-Objects) on a few sample images. These have much more accurate object boundaries than the best variant of EM-Adapt [28], which tends to localize the object well, but produces a ‘blob-like’ segmentation, cf. last four rows in the figure in particular. The first three rows show example images containing multiple object categories. M-CNN recognizes object classes more accurately, e.g., cow in row 5, than EM-Adapt, which confuses cow (shown in green) with horse (magenta). Furthermore, our segmentation results compare favorably with the fully-supervised DeepLab [8] approach (see rows 4-6), highlighting the impact of motion to learn segmentation. There is scope for further improvement, e.g., overcoming the confusion between similar classes in close proximity to each other, as in the challenging case in row 3 for cat vs dog.

Comparison to the state of the art. Table 3 shows our evaluation on the VOC 2012 test set, with our model trained on 20 classes. We performed this by uploading our segmentation results to the evaluation server, as ground truth is not publicly available for the test set. We compare with several state-of-the-art methods with scores taken directly from the publications, except [28] without the post-processing CRF step. This result, shown as ‘[28]’ in the table, is with a model we trained on the VOC augmented dataset. We train M-CNN on all the 20 VOC classes with the model trained (and fine-tuned) on YouTube-Objects and perform a second fine-tuning step together with videos from YouTube-Objects and images from VOC. This achieves 39.8 mean IoU over all the 20 classes, and 49.6 on the 10 classes with video data. This result is significantly better than recent methods using only weak labels, which achieve 25.7 [31], 35.6 [30] and 35.2 [28]. The improvement shown by our M-CNN is more prominent when we consider the average over 10 classes where we use soft motion segmentation cues (and the background), with nearly 10% and 9% boost over [30] and [28] respectively. We also show the evaluation of the model trained on ImageNet-VID in the table.

A few methods have used additional information in the training process, such as the size of objects (+ sz in the table), superpixel segmentation (+ sp), or post-processing steps, e.g., introducing a CRF with pairwise terms learned from fully-annotated data (+ CRF), or even strong or full supervision, such as bounding box (+ bb) or pixel-level

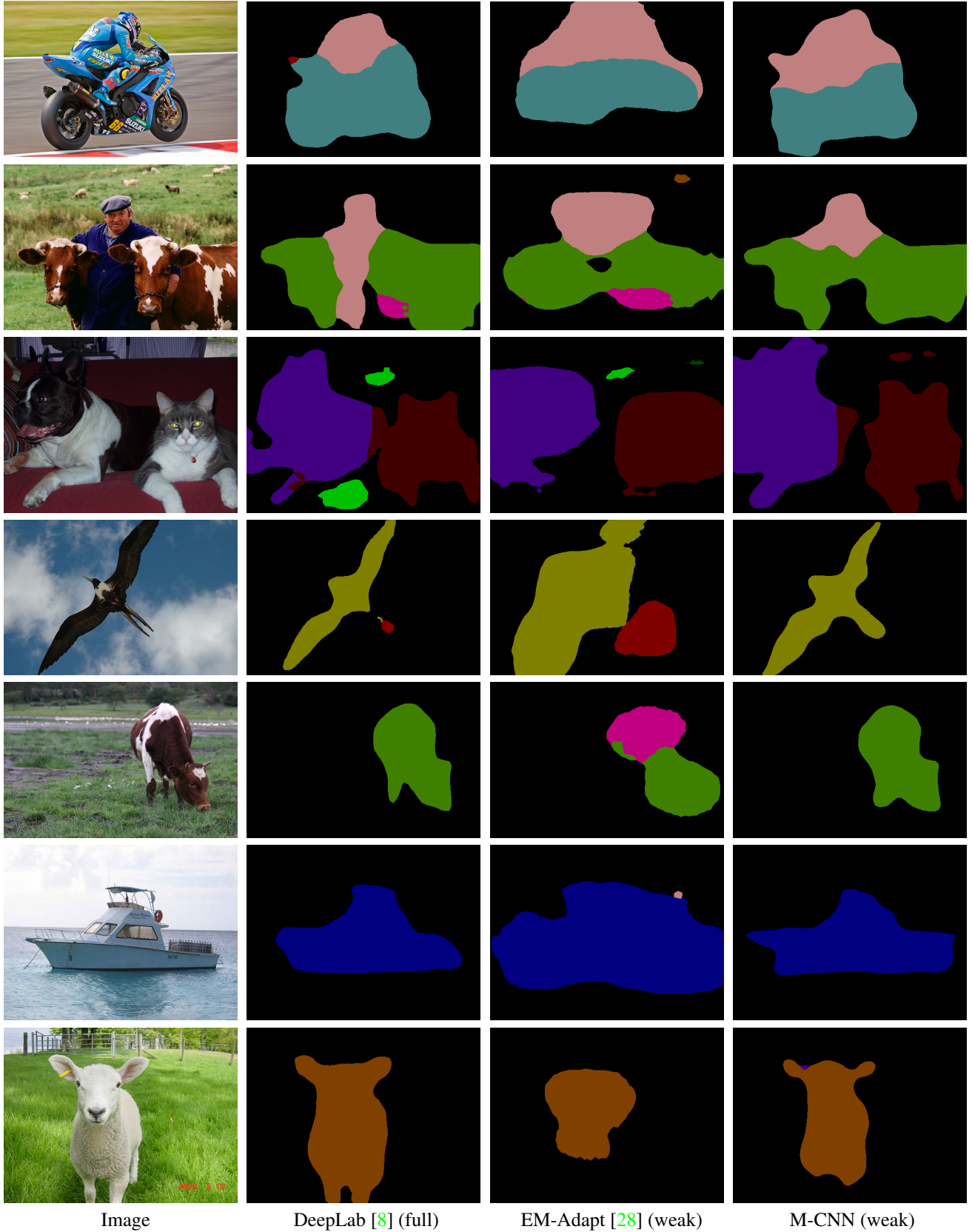


Figure 4. Sample results on the VOC 2012 validation set. Results of fully-supervised DeepLab [8], weakly-supervised EM-Adapt [28] trained on augmented VOC, and our weakly-supervised M-CNN trained on VOC+YouTube-Objects are shown in 2nd, 3rd and 4th columns respectively. (*Best viewed in color.*)

Method	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	Average
<i>Unsupervised</i>											
[5]	53.9	19.6	38.2	37.8	32.2	21.8	27.0	34.7	45.4	37.5	34.8
[29]	65.4	67.3	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1
[21]	55.2	58.7	53.6	72.3	33.1	58.3	52.5	50.8	45.0	19.8	49.9
<i>Weakly supervised</i>											
[33]	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5
[19]	25.1	31.2	27.8	38.5	41.2	28.4	33.9	35.6	23.1	25.0	31.0
[21]	56.5	66.4	58.0	76.8	39.9	69.3	50.4	56.3	53.0	31.0	55.7
M-CNN	76.1	57.7	77.7	68.8	71.6	75.6	87.9	71.9	80.0	52.6	72.0

Table 4. Co-localization performance of M-CNN. We report per class CorLoc scores, and compare with state-of-the-art unsupervised [5, 21, 29] and weakly supervised [19, 21, 33] methods. See text for details.

segmentation (+ seg) annotations. Even though our pure weakly-supervised method is not directly comparable to these approaches, we have included these results in the table for completeness. Nevertheless, M-CNN outperforms some of these methods [28, 32], due to our effective learning scheme. Also from Table 3, the number of training samples used for M-CNN (number of videos shots + number of VOC training images) is significantly lower than those for all the other methods.

4.5. Co-localization

We perform co-localization in the standard setting, where videos contain a common object. Here, we use our M-CNN trained on the YouTube-Objects dataset with 10 categories. We evaluate it on all the frames in YouTube-Objects to obtain prediction scores \mathbf{p}_i for each pixel i . With these scores, we compute a foreground GMM by considering pixels with high predictions for the object category as foreground. A background GMM is also computed in a similar fashion. These form the unary term ψ_i^m in the energy function (2). We then minimize this function with graph cut based inference to compute the binary (object vs background) segmentation labels. Since we estimate segmentations for all the video frames, we do this at the superpixel level [2] to reduce computation cost. We then extract the bounding box enclosing the largest connected component in each frame, and evaluate them following [33]. Quantitative results with this are summarized as per-class and average CorLoc scores in Table 4. We observe that our result outperforms previous state of the art [21] by over 16%. Performing this experiment with ImageNet-VID data we obtain 42.1 on average, in comparison to 37.9 of [29]. ImageNet-VID being a more challenging dataset than YouTube-Objects results in a lower performance for both these methods.

We qualitatively demonstrate the performance of our method on the YouTube-Objects dataset in Figure 5. Our method produces stable results on a variety of categories (third column in the figure). The performance of the mo-

tion segmentation method [29] is also shown for comparison. It is limited by the quality of optical flow and the heuristics used to distinguish foreground from background motion. As a result, it often fails, see second column in the figure.

5. Summary

This paper introduces a novel weakly-supervised learning approach for semantic segmentation, which uses only class labels assigned to videos. It integrates motion cues computed from video as soft constraints into a fully convolutional neural network. Experimental results show that our soft motion constraints can handle noisy motion information and improve significantly over the heuristic size constraints used by state-of-the-art approaches for weakly-supervised semantic segmentation, i.e., by EM-Adapt [28]. We show that our approach outperforms previous state of the art [28, 30] on the PASCAL VOC 2012 image segmentation dataset, thereby overcoming domain-shift issues typically seen when training on video and testing on images. Furthermore, our weakly-supervised method shows excellent results for video co-localization and improves significantly over several recent methods [19, 21, 29].

Acknowledgments. This work was supported in part by the ERC advanced grant ALLEGRO, the MSR-Inria joint project, a Google research award and a Facebook gift. We gratefully acknowledge the support of NVIDIA with the donation of GPUs used for this research.

References

- [1] <http://thoth.inrialpes.fr/research/weakseg>.
- [2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

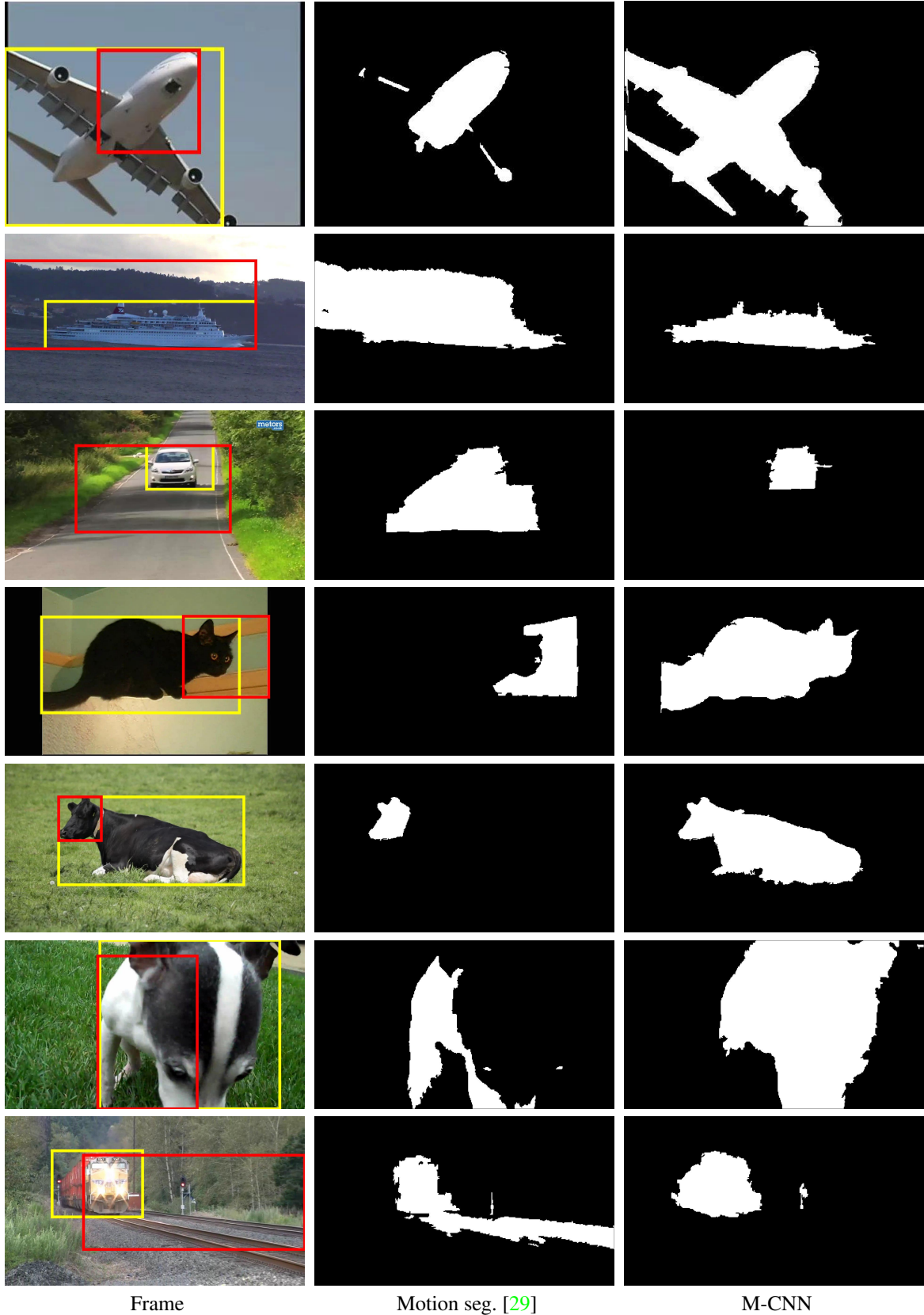


Figure 5. Sample co-localization results on the YouTube-Objects dataset. In the first column the estimated bounding boxes are shown, where yellow corresponds to our result, and red to that of [29]. Segmentations corresponding to [29] and our method are shown in columns 2 and 3 respectively. (*Best viewed in color.*)

- [3] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *ICCV*, 2001.
- [4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.
- [5] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.
- [6] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012.
- [7] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 34(7):1312–1328, 2012.
- [8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.
- [9] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015.
- [10] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *CVPR*, 2013.
- [11] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL training for weakly supervised object localization. In *CVPR*, 2014.
- [12] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014.
- [13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [14] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35(8):1915–1929, 2013.
- [15] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.
- [16] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV*, 2012.
- [17] <http://vision.cs.unc.edu/ilsvrc2015/download-videos-3j16.php#vid>.
- [18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, 2014.
- [19] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In *ECCV*, 2014.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [21] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid. Unsupervised object discovery and tracking in video collections. In *ICCV*, 2015.
- [22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, 1989.
- [23] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*, 2015.
- [24] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.
- [25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [26] A. Monroy and B. Ommer. Beyond bounding-boxes: Learning object shape by model-driven grouping. In *ECCV*, 2012.
- [27] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *CVPR*, 2015.
- [28] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. In *ICCV*, 2015.
- [29] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.
- [30] D. Pathak, P. Krähenbühl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.
- [31] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015.
- [32] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.
- [33] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

- [34] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graphics*, 23(3):309–314, 2004.
- [35] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs. In *CVPR*, 2006.
- [36] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.
- [37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [38] K. D. Tang, R. Sukthankar, J. Yagnik, and F. Li. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013.
- [39] P. Tokmakov, K. Alahari, and C. Schmid. Weakly-supervised semantic segmentation using motion cues. In *ECCV*, 2016.
- [40] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012.
- [41] J. Wu, Y. Zhao, J. Zhu, S. Luo, and Z. Tu. MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *CVPR*, 2014.
- [42] <http://calvin.inf.ed.ac.uk/datasets/youtube-objects-dataset>.
- [43] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.